FP7-212214

Title	**Logging of portal data use, for statistical reporting and evaluation purposes (D5.5)**
Work Package	WP5
Authors	Atle Alvheim
Date	August 2009
Dissemination Level	PU (Public)

Summary/abstract

When data dissemination procedures are automated, both for control and documentation purposes statistics on data use will need to be recorded.

A data access policy requires three types of information:

- In what role does a user access data, what are the rights represented by that role;
- What is the purpose of accessing data, what does the user intend to do;
- Which data resources are accessed?

When users are authenticated and located in a role and have specified a purpose, they are given certain rights of access to data resources. Data resources allow access for specific roles.

A common policy requires definition of user roles, legitimate purposes and access categories.

When data are accessed via the Common CESSDA Portal, information on users and data use is automatically recorded to a central database.

Starting point

If data are published to the Internet, data will be available to anybody. If there are any restrictions on data use, the archives have to develop a system that controls access to protect data against unauthorised use. Further, if there are any requirements for documentation of use, relevant information has to be collected. And last, if data owners request references in material produced on the basis of data, or request reporting back on use of data, a procedure has to ensure that users are informed about the requirements and if necessary is followed up to deliver reports.

The greater transparency of the data offer (online catalogues) forces the data marked to move towards simplification. The integration of the European data catalogues into one single catalogue calls for a clever handling of the data resources that are offered by several archives. As a general principle for a common catalogue, a data resource should only be stored in one place. If a resource is published by more than one data archive, there needs to be an identification system in use that makes it possible to establish if data resources are identical copies. This requires a working versioning system that makes it possible to establish if data resources are non-identical versions of one data resource.

Instead of a Data Trans-border Agreement, CESSDA need a common data access policy, a data Schengen Agreement, a user Registration and Statistics Trans-border Agreement.

Technical framework

The starting point should be a basic data access control architecture that functions broadly enough to implement any specified data access policy. These could be:

Resource-specific rules defined as part of metadata associated with the resource. It is not the intention to introduce restrictions or policies below study level. Neither is it intended at this stage to introduce any payments for data access. And, the system does not contain any disclosure control; a disclosure control system could be regarded as a next step. This requires that data should be anonymised or available for publishing.

For this we require 3 types of information:

- What are the rights or role of the user;
- What does the user intend to do;
- Which data resource is requested.

To establish a manageable policy, it is important that each of these variables represent a clearly defined scale with a limited number of alternatives.

Most archives will have to classify users in groups, define user roles. All members of a role obtain the same rights. Examples could be the role of a researcher conducting academic research, clearly distinguished from the role of a student writing a master thesis. The role becomes the combination of a position and the justification for the work being conducted (motive).

Data resources likewise have to be grouped by access rules. All data resources in the same resource group have the same access rules

Activities combining user roles with resource types could be:

- Access only metadata, which should be free;
- Analyse data online (remote analysis);
- Download data for local use, which is expected to be the typical case.

Remote analysis may be used to identify single data units, if it is possible based on the information in a dataset. However, it is possible to develop disclosure control techniques within the software made available.

User groups

The CESSDA Common Portal is supposed to be open to all. Then we need to know which potential user groups should not be permitted access to the data resources we make available, and we need to know which situations create the disputable problems.

A user role is a combination of position and purpose/motive. A university employed researcher may use data for at least four different reasons: teaching, academic research, private consulting, public consulting. It is not likely that the researcher X has the same rights of access to data resource Y for private consulting as for academic research.

The position of a person is reasonably easy to verify, for objective however there has to be more trust involved.

Position would normally be decided through institutional affiliation. The decisive categories for purposes are the distinctions between non-commercial or commercial activities.

Classification of resources

A CESSDA Common Portal needs to classify resources to establish access rules. A more thorough analysis is required, and the classification of resources to some degree becomes a moving target. Some preliminary work has brought up the following legitimate stakeholders and justifications:

Data inspectorates	Anonymity and data privacy questions
Data owner	Commercial concerns (Payment / market)
	Publicity – References when data used
	Information needs – who is using data for what purpose
	Control – user-control over access (Often other researchers)

Mechanisms

Data access will rarely be answered with a simple "Yes" or "No", but more of some kind of dialogue to collect further information. When all necessary information is collected, there is a definite response. Users therefore have to participate in a dialogue. Positive responses increase user rights.

Registration: If a control unit is to hold information on a user, the user has to register. Until the user is registered somewhere, there are no rights. Either registration is directly in connection with the access system, or the access system works in relation with other registration systems, e.g. institutional systems. Then the big question becomes how and what information may be exchanged.
Authentication: This means that a user, via a password or other mechanism verifies that she is user X. This action usually comes at the start-up of a session if it is necessary to know the user role. For the CESSDA Portal, this action will be triggered by the crossing over from accessing metadata to accessing data. This generates a problem if a user downloads a free resource, then we would not have enough information to link such an action to any user information. It should be investigated if the portal needs to develop an authentication process before any actual data use. So far the answer has been regarded as "no", since the problem only concerns free data resources, and the actual number of downloads may be counted.
Signing pledge of secrecy or similar: This is a verification that the rules of access and secrecy have been read and understood, similar to acceptance of licensing conditions for software. Rules are sent to the user screen the first time there is a need for verification. Responses may either be automated, clicking on a button and response stored in the database, or that the user prints the schema on paper and signs in handwriting, before sending / faxing over a copy. When the user is authorised, rights in the system increase.

Conditions for use: Verification may be obtained by techniques as described above, conditions may be:
Permission: A directly specified permission to use data-resources, usually with a justification. (This could require contacting data owner). This technique requires authentication.
Reporting: Many data owners require reporting, and the importance of this is expected to increase.
Deletion of data after use, no second-hand distribution
Very often it is a condition for access that data are deleted after use. This could be coupled with a time-frame. When time is up, a user could be requested to verify that data have been deleted.

In connection with a system like a CESSDA Common Portal there are logically two conditions that are important to automate as much as possible in a convenient way:

Authorisation: By this it is meant verification of user actions to get access. This could be registration in a database or signing of a sheet of paper. Use of information registered in a database by the user himself functions fairly automatically, but often this has to be coupled with a control or verification process. It depends on what the purpose of information collection is; if it is for the data archive to obtain documentation of activity, or if it is based on data owners' needs to control access. This process may generate delay. Only when

registration processes are authorised and user database updated will the user obtain access. Authorisation should therefore include as few actions as possible.

Repetition: An interactive system differs from a manual system, user expectations are very different. Avoidance of repetition is extremely important. Authorization once should therefore lead to increased rights of access, across categories. If data owners require reporting, that should be organised as self-reporting, maybe strengthened through notification or varieties of follow-up through automatic means, e.g. when accessing the portal next time.

Information collected in a Single Sign-On situation

A single sign-on solution has several justifications:

- Solve problems of multiple passwords required for multiple applications;
- Scaling the account management of multiple applications;
- Security and privacy issues;
- Interoperability with and across organizational boundaries;
- Enabling institutions to define authentication policies, choose technology and build controlled access to resources.

A Single Sign-On solution is basically a dialogue between a data user, a resource provider, a "where are you from" service and an identity provider (often a home institution of the researcher). To study what information such a dialogue may give us, see prototype description under WP12.

A home organisation is fairly important in an automated authentication system. If the researcher has a home organisation, and the home organisation is willing and able (formally and technically) to supply the information we need for our data access policy, then we shift our focus to record information on the user activities.

However, there are users that do not have home organisations, or lack home organisations that have possibilities to support authentication procedures at this level of sophistication, or have home organisations that are not allowed to distribute personal information of this kind. In that case, we have to set up an alternative identity-provider and introduce a registration procedure to get the information we need for our access policies.

Another problem is reporting requirements presented by the data owner. This requirement could be stored with the data resource, so that information on conditions for use is linked to the resource.

Then we get various types of questions:

- Where do we set up and maintain the registration application and where do we store and maintain the database storing the information? This would certainly be a judicial question in addition to a technical.

- How do we validate the information given?

- Do we need to validate the information given? For some types of authorisation: Yes.

- What information/attributes do we have to collect? Only for authorisation or for more purely statistical purposes? Then we would be interested in recording user activities.

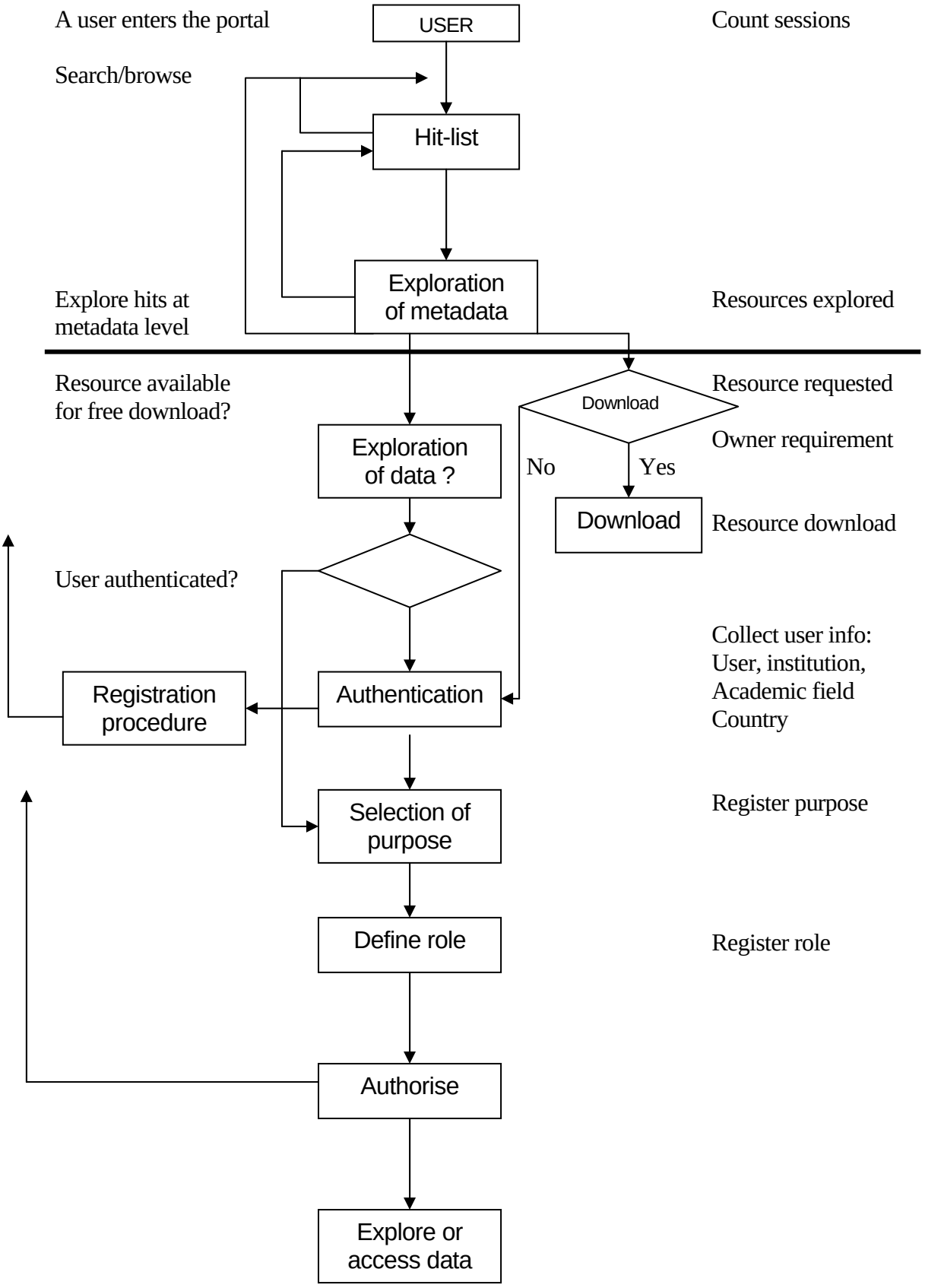
- Technically there would not be any problems connected to designing a database and store such information on portal sessions. But who should be allowed to access the database and download reports?

- As a sequential process, what information do we get?

A user enters the system	Start of session, Count sessions
The user searches for data, and as a response receives a hit list Resources are classified in access categories Resources document owners access conditions	
The user may start another search	Which resource is explored?
The user may explore metadata, then return to hit-list Or start another search	
The user may explore data, or start a download	
Download only available from explore technology	Requested resource
Is the resource freely available for download? If yes, download If no, authenticate user	Which resource is downloaded?
Here a user authentication procedure is initiated	
Via home institution	Register home institution Handle, collect user info
Via other identity provider Can not connect user info for free resources	Collect user info
For authenticated users, what purpose? Select from list (Ask for project title?)	Register purpose
Authenticated user + purpose = Role Role defines user rights in system	Register role
Here is initiated an authorization procedure Role x resource access category decides authorization	Connect user, resource, purpose, owner conditions

The most important information is:

- User, users institution, academic field, country;
- Which resource;
- Purpose of use.



Accessing more than one data repository requires a common federated sign-on.

Illustration of a potential setup to analyse components involved in an access policy

<table>
<tr><td colspan="2">Registration procedure</td><td>→</td><td colspan="4">Authentication procedure control institutional affiliation</td></tr>
<tr><td colspan="2" rowspan="3">|
|
|
|
|
Purpose</td><td rowspan="3"></td><td colspan="2">Academic, according to a list
Defined in database</td><td colspan="2">Non-academic, all other inst.
or lack of institution</td></tr>
<tr><td colspan="4">Project financing</td></tr>
<tr><td>No or financed by own institution</td><td>Research Council or similar, legitimate additional financing for an academic researcher</td><td>Project, for public or private interests outside list of legitimate sources</td><td>Private/ Commercial use of data</td></tr>
<tr><td>Position</td><td>Motive</td><td>Role</td><td colspan="2">Non-Commercial</td><td colspan="2">Commercial</td></tr>
<tr><td>Pupil high-school etc
Student lower degree
Student higher degree</td><td>Some kind of paper/thesis</td><td>Student</td><td colspan="2">Free access,
status confirmed by tutor</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Doctorial student
Academic researcher</td><td>Teaching
Doctorial thesis
Academic research</td><td>Academic researcher</td><td colspan="2">Free access</td></tr>
<tr><td>Project-related/ private researcher

Private-/commercial consultant/researcher</td><td>Projects
public
private
Other research
Journalistic work</td><td>Non-academic researcher</td><td colspan="2"></td><td colspan="2">Very often specific conditions,
Contact data owner, payment, etc

manual procedures required</td></tr>
</table>

All archives require an explicit access policy.

The portal requires a harmonised access policy to be able to develop a federated Single Sign On.

The harmonised policy has to define user roles and grouping of conditions for dataset access.

The table above is an indication of how we could go about analysing the relationship between variables. These are the complicated questions. To develop a central database to collect the information is a simple task.